STAT100 Problem Set 3: Exploratory Data Analysis 2

You need to submit a Word document or PDF for this assignment. Make sure you do the following:

1. **Upload only one file/document in ELMS for Problem Set 3.**
2. **Include your name in the document in the upper left-hand corner. Under your name, write STAT 100 and your section number. Write Problem Set 3 centered on the page.**
3. **Number and letter your answers to the questions accordingly.**
4. **Carefully read all problems and follow all instructions.**
5. **Upload the assignment in ELMS before the deadline of Sunday 2/27 at 11:59 PM otherwise it is considered late. Make sure you save your document on your computer or email it to yourself so that you keep an electronic copy.**

STAT100 Problem Sets need to be completed by students in RStudio. Students should refer to the Tutorial for Problem Set 3 as they are working on this problem set. All Tutorials for Problem Sets can be found in the STAT100 ELMS course under Modules.

For this assignment you need to use the *Course_Data_Set.RData* file, which includes data collected from students enrolled in introductory statistics courses in various colleges and universities in MD. The population of interest is all students enrolled in introductory statistics courses in the US.

Problem Set 3 has two questions worth 25 points. Read each question carefully and follow all instructions. **Please follow these instructions for providing your responses:**

- **For #1.a., 1.b., 2.a. and 2.d., you need to provide the R code that you use to generate the output or data display and provide the output as displayed in the RStudio Console or the data display as shown in the Zoom window in RStudio.**
- **For all other questions, you should type your responses directly in the document you submit for Problem Set 3. You should NOT provide R code or RStudio output for all other questions.**

1. (10 points) **Open the** *Course_Data_Set.RData* **file in RStudio.** You need to generate descriptive statistics for variables and answer questions related to the distribution of two variables.

    a. For the *Height_inches* variable, use the summary command to generate descriptive statistics. The *Height_inches* variable measures the height (in inches) of the students in the Course Data Set. In the document you upload for this assignment, include:
        i. the R code you used to generate the descriptive statistics, including any comments **(IMPORTANT INSTRUCTIONS: you MUST include your name in a comment line of the R code for #1.a.)**
        ii. an image of the descriptive statistics output in RStudio

    b. For the *HS_GPA* variable, use the summary command to generate descriptive statistics. The *HS_GPA* variable measures the high school grade point average (GPA) of the students in the Course Data Set. In the document you upload for this assignment, include:
        i. the R code you used to generate the descriptive statistics, including any comments **(IMPORTANT INSTRUCTIONS: you MUST include your name in a comment line of the R code for #1.b.)**
        ii. an image of the descriptive statistics output in RStudio

    c. Using the 1.5*(IQR) criterion for outliers, for the distribution of students enrolled in introductory statistics courses, would a height of 80 inches be considered an outlier? *Show all work*.

    d. Using the 1.5*(IQR) criterion for outliers, for the distribution of students enrolled in introductory statistics courses, would a high school GPA of 1.85 be considered an outlier? *Show all work*.

2. (15 points) **Open the** *Course_Data_Set.RData* **file in RStudio.** You need to create histograms for variables and answer questions related to the distribution of the variables.

    a. For the *Height_inches* variable, create two histograms (one with the default number of bins/intervals, and one with 6 bins/intervals). Your histogram should have proper axis labels. In the document you upload for this assignment, include:
        i. the R code you used to generate the histograms, including any comments
        ii. separate images of the histograms from the Plots/Zoom section of RStudio. **IMPORTANT INSTRUCTIONS: the titles of the histograms MUST include your name.** The title of the histogram with default bins MUST read “Distribution of Student Height with default bins, created by *FirstName LastName*”. For example, if Prof. Griffin created the histogram, the title would read “Distribution of Student Height with default bins, created by Matt Griffin”. The title of the histogram with 6 bins MUST read “Distribution of Student Height with 6 bins, created by *FirstName LastName*”.

b. Based on the histograms, describe the distribution of student height (shape, center, spread, outliers). Does your interpretation of the distribution of student height differ depending on which histogram you examine (number of bins/intervals)?

c. Using only one of the histograms for student height, within which interval does the median height fall? **In your response to this question, do NOT refer to the descriptive statistics generated for #1.a.** Provide the endpoints of the interval, indicate which histogram you used (number of bins/intervals), and explain how you arrived at your answer.

d. For the *HS_GPA* variable, create two histograms (one with the default number of bins/intervals, and one with 8 bins/intervals). In the document you upload for this assignment, include:

    i. the R code you used to generate the histograms, including any comments

    ii. separate images of the histograms from the Plots/Zoom section of RStudio. **IMPORTANT INSTRUCTIONS: the titles of the histograms MUST include your name.** The title of the histogram with default bins MUST read "Distribution of High School GPA with default bins, created by *FirstName LastName*". For example, if Prof. Griffin created the histogram, the title would read "Distribution of High School GPA with default bins, created by Matt Griffin". The title of the histogram with 8 bins MUST read "Distribution of High School GPA with 8 bins, created by *FirstName LastName*".

e. Based on the histograms, describe the distribution of high school GPA (shape, center, spread, outliers). Does your interpretation of the distribution of high school GPA differ depending on which histogram you examine (number of bins/intervals)?

f. Using only one of the histograms for high school GPA, within which interval does the median high school GPA fall? **In your response to this question, do NOT refer to the descriptive statistics generated for #1.b.** Provide the endpoints of the interval, indicate which histogram you used (number of bins/intervals), and explain how you arrived at your answer.